Learning and Equilibrium

Learning and Equilibrium

Drew Fudenberg* and David K. Levine**

First version: June 9, 2008

This version: August 30, 2008

* Department of Economics, Harvard University, Cambridge MA, dfudenberg@harvard.edu

** Department of Economics, Washington University of St. Louis, St. Louis, MI, david@dklevine.com

Acknowledgments: We are grateful to Sergiu Hart, Josef Hofbauer, Bill Sandholm, Satoru Takahashi, Yuichi Yamamoto, and Peyton Young for helpful comments, and to NSF grants SES-03-14713 and SES- 06-646816 for financial support.

Keywords: non-equilibrium dynamics, bounded rationality, Nash equilibrium, self-confirming equilibrium

Abstract: The theory of learning in games studies how, which and what kind of equilibria might arise as a consequence of a long-run non-equilibrium process of learning, adaptation and/or imitation. If agents' strategies are completely observed at the end of each round, and agents are randomly matched with a series of anonymous opponents, fairly simple rules perform well in terms of the agent's worst-case payoffs, and also guarantee that any steady state of the system must correspond to an equilibrium. If (as in extensive-form games) players do not observe the strategies chosen by their opponents, then learning is consistent with steady states that are not Nash equilibria because players can maintain incorrect beliefs about off-path play. Beliefs can also be incorrect due to cognitive limitations and systematic inferential errors.

Table of Contents

. Introduction

This article reviews the literature on non-equilibrium learning in games, with a focus on work too recent to have been included in our book *The Theory of Learning in Games* (1998). Due to space constraints, the article is more limited in scope, with a focus on models of how individual agents learn, and less discussion of evolutionary models and models of myopic adjustment.[1]

Much of the modern economics literature is based on the analysis of the equilibria of various games, so the issue of when and why to expect observed play to resemble an equilibrium is of primary importance. Rationality (as defined for example by Savage (1954)) does not imply that the outcome of a game must be a Nash equilibrium, and neither does common knowledge that players are rational, as equilibrium requires all players to coordinate on the same equilibrium. However, game theory experiments show that the outcome after multiple rounds of play is often much closer to equilibrium predictions than play in the initial round, which supports the idea that equilibrium arises as a result of players learning from experience. The theory of learning in games formalizes this idea, and examines how, which and what kind of equilibrium might arise as a consequence of a long-run non-equilibrium process of learning, adaptation and/or imitation. Our preferred interpretation and motivation for this work is not that the agents are "trying to reach Nash equilibrium," but rather that they are trying to maximize their own payoff while simultaneously learning about the play of other agents. The question is

[1] For recent surveys of evolutionary game theory see Hofbauer & Sigmund (2003) and Sandholm (2009).

then when self-interested learning and adaptation will result in some sort of equilibrium behavior.

It is not satisfactory to explain convergence to equilibrium in a given game by assuming an equilibrium of some larger dynamic game in which player choose adjustment or learning rules knowing the rules of the other agents. For this reason, in the models we survey there are typically some players whose adjustment rule is not a best response to the adjustment rules of the others, and so it is not a relevant criticism to say that some player's adjustment rule is sub-optimal. Instead, the literature has developed other criteria for the plausibility of learning rules, such as there not being relatively obvious and simple alternatives that would be better.

The simplest setting in which to study learning is one in which agents' strategies are completely observed at the end of each round, and agents are randomly matched with a series of anonymous opponents, so that the agents have no impact on what they observe. We discuss these sorts of models in section 2. Section 3 discusses learning in extensive-form games, where it is natural to assume that players do not observe the strategies chosen by their opponents but (at most) the sequence of actions that were played. That section also discusses models of some frictions that may interfere with learning, such as computational limits or other causes of systematic inferential errors.

2. Learning in Strategic Form Games

In this section we consider settings where players do not need to experiment to learn. Throughout this section we assume that players know their own payoffs and see the action employed by their opponent in each period of a simultaneous move game; the case

in which players do not know their own payoffs is discussed in section 3 when we examine extensive form games.

The experimental data on how agents learn in games is noisy,[2] so the theoretical literature has relied on the idea that people are likely to use rules that perform well in situations of interest, and also on the idea that rules should strike a balance between performance and complexity. In particular, simple rules perform well in simple environments, while a rule needs more complexity to do well when larger and more complex environments are considered.

Section 2A discusses work on fictitious play and stochastic fictitious play. These models are relatively simple, and have the interpretation as the play of a Bayesian agent who believes he is facing a stationary environment. These models also "perform well" when the environment (in this case, the sequence of opponent's plays) is indeed stationary or at least approximately so. The simplicity of this model gives it some descriptive appeal, and also makes it relatively easy to analyze using the techniques of stochastic approximation. However, with these learning rules play only converges to Nash equilibrium in some classes of games, and when play does not converge the environment is not stationary and the players' rules may perform poorly. Section 2B discusses various notions of "good asymptotic performance," starting from Hannan-consistency, which means doing well in stationary environments, and moving on to stronger conditions that ensure good performance in more general settings. Under calibration, which is the strongest of these concepts, play converges globally to the set of correlated equilibria.

[2] There is an extensive literature that tries to identify and estimate the learning rules used by subjects in game theory experiments, see for example by Cheung & Friedman (1997), Erev & Roth (1998), and Camerer & Ho (1999). However, Salmon (2001) shows that experimental data has little power in

This leads us to discuss the related question of whether these more sophisticated learning rules imply that play always converges to Nash equilibrium. Section 2C discusses models where players act as if they do not know the payoff matrix, including reinforcement learning models adapted from the psychology literature and models of imitation. It also discusses the interpretation of stochastic fictitious play as reinforcement learning.

2A. Fictitious play and stochastic Fictitious Play

Fictitious play (FP) and stochastic fictitious play (SFP) are simple stylized models of learning. They apply to settings where the agents repeatedly play a fixed strategic-form game. The agent knows the strategy spaces and her own payoff function, and observes the strategy played by her opponent in each round. The agent acts as if she is facing a stationary but unknown (exchangeable) distribution of opponents' strategies, so she takes the distribution of opponents' play as exogenous. To explain this "strategic myopia," Fudenberg & Kreps (1993) appealed to a "large population model" with many "agents" in each "player role." Perhaps the best example of this is the model of *anonymous random matching:* Each period all agents are matched to play the game, and are told only play in their own match. Agents are unlikely to play their current opponent again for a long time, even unlikely to play anyone who played anyone who played her. So if the population size is large enough compared to the discount factor, it is not worth sacrificing current payoff to influence this opponent's future play.

---

discriminating between alternative learning models; this is supported by Wilcox (2006)'s finding that the assumption of a representative agent can drive some of the conclusions of this literature.

In FP, players act as if they are Bayesians; they believe that the opponents' play corresponds to draws from some fixed but unknown mixed strategy,[3] and belief updating has a special simple form: Player i has an exogenous initial weight function $\kappa_0^i : S^{-i} \to \Re_+$, where S^{-i} is the space of opponents' strategies.[4] This weight is updated by adding 1 to the weight of each opponent strategy each time it is played, so that

$$\kappa_t^i(s^{-i}) = \kappa_{t-1}^i(s^{-i}) + \begin{cases} 1 \text{ if } s_{t-1}^{-i} = s^{-i} \\ 0 \text{ if } s_{t-1}^{-i} \neq s^{-i} \end{cases}.$$

The probability that player i assigns to player $-i$ playing s^{-i} at date t is given by

$$\gamma_t^i(s^{-i}) = \frac{\kappa_t^i(s^{-i})}{\sum_{\tilde{s}^{-i} \in S^{-i}} \kappa_t^i(\tilde{s}^{-i})}.$$

Fictitious play is any behavior rule that assigns actions to histories by first computing γ_t^i and then picking any action in $BR^i(\gamma_t^i)$. As noted by Fudenberg & Kreps, this update rule corresponds to Bayesian inference when player i believes that the distribution of opponents' strategies corresponds to a sequence of i.i.d. multinomial random variables with a fixed but unknown distribution, and player i's prior beliefs over that unknown distribution take the form of a Dirichlet distribution. While this form of the prior simplifies the formula for updating beliefs, it is not important for the qualitative results; what is important is the implicit assumption that the player treats the environment

[3] Note that with the large population interpretation, this belief does not require that any agent actually randomizes her play. The belief that opponents play, say, (2/3 L, 1/3 R) is consistent with a state where 2/3 of the opponents always play L and 1/3 always play R.

[4] For expositional simplicity, we focus on 2-player games here. Fudenberg & Kreps (1993) discussed the conceptual issues in extending FP to games with three or more players.

as stationary. This ensures that the assessments will converge to the marginal empirical distributions.

If all agents use FP then the actual environment is not stationary unless they start at steady state, so agents have the wrong model of the world. But stationarity is a reasonable first hypothesis in many situations. This is not to say, however, they we expect it to be maintained by agents when it obviously fails, as for example, when fictitious play generates high frequency cycles. Consider following example from Fudenberg & Kreps

	A	B
A	0,0	1,1
B	1,1	0,0

Suppose there is 1 agent per side, both use FP with initial weights $(1,\sqrt{2})$ for each player. In the 1st period, both players think the other will play B, so both play A. The next period the weights are $(2,\sqrt{2})$ and both play B; the outcome is the alternating sequence ((B.,B),(A,A),(B,B),….). In FP players only randomize when exactly indifferent, so typically per-period play cannot converge to a mixed-strategy Nash equilibrium, but it is possible for the empirical frequencies of each player's choices to converge to a mixed Nash equilibrium, as they do in this example. However, the realized play is always on the diagonal, so both players receive payoff 0 in every period and the empirical distribution on action profiles does not equal the product of the two marginal distributions. This does

not seem a very satisfactory notion of "converging to an equilibrium," and it shows the drawbacks of identifying a cycle with its average.[5]

Stochastic Fictitious Play

In the process of "stochastic fictitious play" or SFP, players form beliefs as in FP but choose actions according to a stochastic best response function. One explanation for the randomness is that it reflects payoff shocks as in Harsanyi's (1973) purification theorem. Here the payoff to each player or agent i is perturbed by i.i.d. random shocks η_t^i that are private information to that agent, and in each period each agent chooses a rule mapping his type (realized payoff) to his strategy. For each distribution $\sigma^{-i} \in \Delta(S^{-i}) \equiv \Sigma^{-i}$ over the actions of *i*'s opponents, define player *i*'s *best-response distribution* (or smooth best response function)

$$\overline{BR}^i(\sigma^{-i})(s^i) = \text{Prob}[\eta^i \text{ s.t. } s^i \text{ is a best response to } \sigma^{-i}].$$

Any opponent's play σ^{-i} induces a unique best response for almost every type, so when the distribution of types is absolutely continuous with respect to Lebesgue measure, the best-response distribution is indeed a function, and moreover it is continuous. For example, the *logit (or logistic) best response is*

$$\overline{BR}^i(\sigma^{-i})(s^i) = \frac{\exp(\beta u(s^i, \sigma^{-i})}{\sum_{s'^i} \exp(\beta u(s'^i, \sigma^{-i})}.$$

[5] Historically FP was viewed as a thought process by which players might compute and perhaps coordinate on a Nash equilibrium without actually playing the game (hence "fictitious.") From this perspective, convergence to a limit cycle was not problematic, and the early papers focused on finding games in which the time average of FP converges. When it does converge, the resulting pair of marginal distributions must be a Nash equilibrium.

When β is large this approximates the exact best response correspondence. Fudenberg & Kreps (1993) called the intersection of these functions a "Nash distribution," because it corresponds to the Nash equilibrium of the static Bayesian game corresponding to the payoff shocks; as β goes to infinity the Nash distributions converge to the Nash equilibrium of the complete-information game.[6]

As compared to FP, SFP has several advantages: It allows a more satisfactory explanation for convergence to mixed-strategy equilibria in fictitious play-like models. For example, in matching pennies the per-period play can actually converge to the mixed strategy equilibrium. In addition, SFP avoids the discontinuity inherent in standard fictitious play, where a small change in the data can lead to an abrupt change in behavior. With SFP, if beliefs converge, play does too. Finally, as we discuss in the next section, there is a (non-Bayesian) sense in which stochastic rules perform better than deterministic ones: stochastic FP is "universally consistent" (or "Hannan-consistent") in the sense that its time average payoff is at least as good as maximizing against the time-average of opponents' play, which is not true for exact FP.

For the analysis to follow, the source of smooth best response function is unimportant. It is convenient to think of it as having been derived from the maximization of a perturbed deterministic payoff function that penalizes pure actions (as opposed the stochastic perturbations in the Harsanyi approach). Specifically, if v^i is a smooth, strictly differentiable, concave function on the interior of Σ^i whose gradient becomes infinite at the boundary, then $\arg\max_{\sigma^i} u^i(\sigma^i,\sigma^{-i}) + \beta^{-1}v^i(\sigma^i)$ is a smooth best response function

[6] Note that not all Nash equilibria can approached in this way, think for example of Nash equilibria in weakly dominated strategies. Following McKelvey & Palfrey (1995), Nash distributions have become

that assigns positive probability to each of i's pure strategies; the logit best response corresponds to $v^i(\sigma^i) = \sum_{s^i} -\sigma^i(s^i)\log\sigma^i(s^i)$. It has been known for a long time that the logit model also arises from a random-payoff model where payoffs have the extreme-value distribution; Hofbauer & Sandholm (2002) extended this. They showed that if the smooth best responses are continuously differentiable and are derived from a "simplified Harsanyi model" where the random types have strictly positive density everywhere, then they can be generated from an "admissible" deterministic perturbation. [7]

Now we consider systems of agents, all of whom use SFP. The technical insight here is that the methods of "stochastic approximation" apply, so that the asymptotic properties of these stochastic, discrete-time systems can be understood by reference to a limiting continuous time deterministic dynamical system. There are many versions of the stochastic approximation result in the literature. The following version from Benaim (1999) is general enough for the current literature on SFP: Consider the discrete time process on a nonempty convex subset X of R^m defined by the recursion $x_{n+1} - x_n = (1/(n+1))\left[F(x_n) + U_n + b_n\right]$, and the corresponding continuous time semi-flow Φ induced by the system of ordinary differential equations $dx(t)/dt = F(x(t))$, where the U_n are mean-0, bounded-variance error terms, and

---

known in the experimental literature as a quantal response equilibrium, and the logistic smoothed best response as the quantal best response.

[7] A key step is the observation that the derivative of the smooth best response is symmetric, and the off-diagonal terms are negative: a higher payoff shock on *i*'s first pure strategy lowers the probability of every other pure strategy. This means the smooth best response function has a convex potential function: a function *W* (representing maximized expected utility) such that the vector of choice probabilities is the gradient of the potential, analogous to the indirect utility function in demand analysis. Hofbauer & Sandholm then show how to use the Legendre transform of the potential function to back out the disturbance function. Note that the converse of the theorem is not true: some functions obtained by maximizing a deterministic perturbed payoff function cannot be obtained with privately observed payoff shocks, and indeed Harsanyi had a counterexample.

$b_n \to 0$. Under additional technical conditions,[8] there is probability 1 that every ω-limit[9] of the discrete-time stochastic process lies in a set that is internally chain-transitive for Φ.[10] (It is important to note that the stochastic terms do not need to be independent or even exchangeable.)

Benaim & Hirsch (1999) applied stochastic approximation to the analysis of SFP in two-player games, with a single agent in the role of player 1 and a second single agent in the role of player 2. The discrete-time system is then

$$\begin{aligned}\theta_{1,n+1} - \theta_{1,n} &= (1/(n+1))[\overline{BR}_1(\theta_{2,n}) - \theta_{1,n} + U_{1,n} + b_{1,n}] \\ \theta_{2,n+1} - \theta_{2,n} &= (1/(n+1))[\overline{BR}_2(\theta_{1,n}) - \theta_{2,n} + U_{2,n} + b_{2,n}]\end{aligned},$$

where $\theta_{i,n}$ is player j's beliefs about the play of player i, the $U_{i,n}$ are the mean-zero error terms, and the $b_{i,n}$ are asymptotically vanishing error terms that accounts for the difference between the player j's beliefs and the empirical distribution of i's play.[11]

They then used stochastic approximation to relate the asymptotic behavior of the system to that of the deterministic system

$$\dot{\theta}_1 = BR_1(\theta_2) - \theta_1, \dot{\theta}_2 = BR_2(\theta_1) - \theta_2 .$$

They also provided a similar result for games with more than two players, still with one agent in each population. Note that the rest points of this system are exactly the

---

[8] Measurability of the stochastic terms, integrability of the semi-flow, and pre-compactness of the x_n.

[9] The ω-limit set of a sample path $\{\theta_n\}$ is the set of long-run outcomes: y is in the ω-limit set if there is an increasing sequence of periods $\{n_k\}$ such that $\theta_{n_k} \to y$ as $n_k \to \infty$.

[10] These are sets that are compact, invariant, and do not contain a proper attractor.

[11] Benaim & Hirsch simplified by ignoring the prior weights so that beliefs are identified with the empirical distributions.

equilibrium distributions. Thus stochastic approximation says roughly that SFP cannot converge to a linearly unstable Nash distribution, and that it has to converge to one of the system's internally chain transitive sets.

Of course, this leaves open the issue of determining the chain transitive sets for various classes of games. Fudenberg & Kreps (1993) established global convergence to a Nash distribution in 2x2 games with a unique mixed-strategy equilibrium; Benaim & Hirsch (1999) provided a simpler proof of this, and established that SFP converges to a stable, approximately pure Nash distribution in 2x2 games with two pure strategy equilibria; they also showed that SFP does not converge in Jordan's (1993) three-player matching pennies game. Hofbauer & Sandholm (2002) used the relationship between smooth best responses and deterministic payoff perturbations to construct Lyapunov function for SFP in zero-sum games and potential games (Monderer & Shapley (1996)) and hence prove (under mild additional conditions) that SFP converges to a steady state of the continuous time system. Hofbauer and Sandholm derived similar results for a one-population version of SFP, where two agents per period are drawn to play a symmetric game, and the outcome of their play is observed by all agents; this system has the advantage of providing an explanation for the "strategic myopia" assumed in SFP.

Ellison & Fudenberg (2000) studied (Unitary) in 3×3 games, in cases where smoothing arises from a sequence of Harsanyi-like stochastic perturbations, with the "size" of the perturbation going to zero. They found that there are many games in which whether a purified version of the totally mixed equilibrium is locally stable depends on the specific distribution of the payoff perturbations, and that there are some games for which no "purifying sequence" is stable. Sandholm (2007) re-examined the stability of

purified equilibria under (Unitary); he gave general conditions for stability and instability of equilibrium, and shows that there is always at least one stable purification of any Nash equilibrium when a larger collection of purifying sequences is allowed. Hofbauer & Hopkins (2005) proved convergence of (Unitary) in all two-player games that can be rescaled to be zero-sum, and in two-player games that can be rescaled to be partnerships. They also showed that isolated interior equilibria of all generic symmetric games are linearly unstable for all small symmetric perturbations of the best response correspondence, where a "symmetric perturbation" means that the two players have the same smoothed best response functions. This instability result applies in particular to symmetric versions of the famous example of Shapley (1964), and to non-constant-sum variations of the game "rock-scissors-paper."[12] The overall conclusion seems to be fairly optimistic about convergence in some classes of games, and pessimistic in others. For the most part, the above papers motivated (Unitary) as describing the long-run outcome of SFP; but Ely & Sandholm (2005) showed that (Unitary) also described the evolution of the population aggregates in their model of Bayesian population games.

Fudenberg & Takahashi (2007) studied "heterogeneous" versions of SFP, with many agents in each player role, and each agent only observing the outcome of their own match. The bulk of their analysis assumes that all agents in a given population have the same smooth best response function.[13] In the case where there are separate populations of "player 1's" and "player 2's," and all agents play every period, the standard results

---

[12] The constant-sum case is one of the non-generic games where the equilibrium is stable.

[13] The perturbations used to generate smoothed best responses may also be heterogeneous. Once this is allowed, the beliefs of the different agents can remain slightly different, even in the limit, but a continuity argument shows that this has little impact when the perturbations are small.

extend without additional conditions. Intuitively, since all agents in population 1 are observing draws at the same frequency from a common (possibly time varying) distribution, they will eventually have the same beliefs. Consequently, it seems natural that the set of asymptotic outcomes should be the same as in a system with one agent per population. Similar results obtain in a model with "personal clocks," where a single pair of agents is selected to play each day, with each pair having a possibly different probability of being selected, provided that (a) the population is sufficiently large compared to the Lipschitz constant of the best-response functions, and (b) the matching probabilities of various agents are not "too different." Under these conditions, although different agents observe slightly different distributions, their play is sufficiently similar that their beliefs are the same in the long run. While this provide some support for results derived from (Unitary), the condition on the matching probabilities is fairly strong, and rules out some natural cases such as interacting only with neighbors; the asymptotics of SFP in these cases is an open question.

Benäim et al. (2007) extended stochastic approximation analysis from SFP to "weighted stochastic FP" in which agents give geometrically less weight to older observations. Roughly speaking, weighted smooth FP with weights converging to 1 gives the same trajectories and limit sets as SFP; the difference is in the speed of motion and hence in whether the empirical distribution converges. They considered two related models, both with a single population playing a symmetric game, unitary beliefs, and a common smooth best response function. In one model, there is a continuum population, all agents are matched each period, and the aggregate outcome X_t is announced at the end of period t. The aggregate common belief then evolves according to

$x_{t+1} = (1-\gamma_t)x_t + \gamma_t X_t$, where γ_t is the step size; because of the continuum of agents, this is a deterministic system. In the second model, one pair of agents is drawn to play each period, and a single player's realized action is publicly announced, all players update according to $x_{t+1} = (1-\gamma_t)x_t + \gamma_t X_t$ where X_t is the action announced at period t. (Standard SFP has step size $\gamma_t = 1/(t+1)$; this is what makes the system "slow down" and leads to stochastic approximation results; it is also why play can cycle too slowly for time averages to exist.)

Consider the system where only one pair plays at a time. This system is *ergodic:* It has a unique invariant distribution, and the time average of play converges to that distribution from any initial conditions.[14] To determine what this invariant distribution is, Benäim et al. focus on the case of weights γ near 0, where the tools of stochastic approximation can be of use. Specifically, they related the invariant distribution to the Birkhoff center[15] of the continuous-time dynamics that stochastic approximation associates with SFP. Specifically, let ν_δ denote the invariant distribution for weighting $\gamma_t = 1-\delta$, and let ν_1 be an accumulation point of ν_δ as $\delta \to 1$. Benaïm et al show that ν_1 is contained in the Birkhoff center of the flow of the smooth best response dynamic. They used this, along with other results, to conclude that if the game payoff matrix is positive definite in the sense that $\lambda^T A\lambda > 0$ for all non-zero vectors λ that sum to 0, if the game has a unique and fully mixed equilibrium x^*, and if the smooth best response

[14] This follows from results of Norman (1968). It is enough to show that the system is "distance diminishing" – the distance between two states goes down after any observation – and that from any state there is positive probability of getting arbitrarily close to the state (1,0,0,…).

[15] The Birkhoff center of a flow is the closure of the set of points *x* such that *x* is contained in the ω-limit from *x;* it is contained in the union of the internally chain transitive sets.

function has the logit form with sufficiently large parameter β, then the limit invariant distribution ν_1 assigns probability 0 to any Nash distribution that is near x^*. This shows that in this game the weighted SFP does not converge to the unique equilibrium. Moreover, under some additional conditions the iterated limit $\beta \to \infty, \gamma \to 0$ of the average play is, roughly speaking, the same cycle that would be observed in the deterministic system.

To help motivate their results, Benäim et al. referred to an experiment of Morgan et al. (2006). The game's equilibria are unstable under SFP, but the aggregate (over time and agents) play looks "remarkably close" to NE, which is consistent with the paper's prediction of a stable cycle. As the authors pointed out, the information decay that gives the best fit on experimental data is typically not that close to 0, and simply having a lower parameter β in unweighted SFP improves the fit as well. As evidence against the unweighted rule, Benaïm et al. note that the experimenters report some evidence of auto-correlation in play; other experiments starting with Cheung & Friedman (1997) have also reported evidence that agents discount older observations. It would be interesting to see how the autocorrelation in the experiments compares with the autocorrelation predicted by weighed SFP, and whether there the subjects were aware of these cycles.

2B. Asymptotic Performance and Global Convergence

SFP treats observations in all periods identically, so it implicitly assumes that the players view the data as exchangeable. It turns out that SFP guarantees that players do at least as well as maximizing against the time average of play, so that when the environment is indeed exchangeable the learning rule "performs well." However, SFP does not require that players identify trends or cycles, which motivates the consideration

of more sophisticated learning rules that perform well in a wider range of settings, This in turn leads to the question of how to assess the performance of various learning rules.

From the viewpoint of economic theory it is tempting to focus on Bayesian learning procedures, but these procedures do not have good properties against possibilities that have zero prior probability (Freedman, 1965). Unfortunately, any prior over infinite histories must assign probability zero to "very large" collections of possibilities.[16] Worse, in interacting with equally sophisticated (or more sophisticated) players, the interaction between the players may force play of opponents to have characteristics that were *a priori* thought to be impossible,[17] which leads us to consider non-Bayesian optimality conditions of various sorts.

Since FP and SFP only tracks frequencies, and not information relevant to identifying cycles or other temporal patterns, there is no reason to expect them to do well except with respect to frequencies, so one relevant non-Bayesian criterion is to get (nearly) as much utility as if the frequencies are known in advance, uniformly over all possible probability laws over observations. If the time average of utility generated by the learning rules attains this goal asymptotically, we say that it is "universally consistent" or "Hannan consistent." The existence of universally consistent learning rules was first proved by Hannan (1957) and Blackwell (1956). A variant of this result was rediscovered in the computer science literature by Banos (1968) and Megiddo (1980), who showed that

[16] If each period has only two possible outcomes, the set of histories is the same as the set of binary numbers between 0 and 1. Consider on the unit interval the set consisting of a ball around each rational point, where the radius of the k th ball is r/k^2. This is big in the sense that it is open and dense, but when r is small the set has small Lebesgue measure. See Stinchcombe (2005) for an analysis using more sophisticated topological definitions of what it means for a set to be small.

[17] Kalai and Lehrer [1993] rule this out by an assumption that requires a fixed-point-like consistency in the players' prior beliefs. Nachbar [1997] shows that "*a priori* impossible" play is unavoidable when the priors are required to be independent of the payoff functions in the game.

there are rules that guarantee a long run average payoff of at least the minmax. The existence of universally consistent rules follows also from the Foster & Vohra's (1997) result on the existence of universally calibrated rules that we discuss below. Notice that universal consistency says that in matching pennies, if the other player plays heads in odd period and tails in even periods, "good performance" is to win half the time, even though it would be possible to always win. This is reasonable, as it would only make sense to adopt "always win" as the benchmark for learning rules that had the ability to identify cycles.

To prove the existence of universally consistent rules, Blackwell (1956b) (discussed in Luce & Raiffa (1957)) used the concept of approachability that was introduced in Blackwell (1956a). Subsequently Hart & Mas-Colell used approachability in a different way to construct a family of universally consistent rules. Benaim et al. (2006) further refine this approach, using stochastic approximation results for differential inclusions. For SFP, Fudenberg & Levine (1995) used a stochastic approximation argument applied to the difference between the realized payoff and the "consistency" benchmark, similar in spirit the original proof of Hannan; subsequently Fudenberg & Levine (1999) used a calculation based on the assumption that the smooth best response functions are derived from maximizing an perturbed deterministic payoff function, and so have symmetric cross partials.[18]

The Fudenberg & Kreps example shows that FP is not universally consistent. However, Fudenberg & Levine (1995) and Monderer et al. (1997) showed that when FP fails to be consistent it must result in the player employing the rule frequently switching

back and forth between his strategies. Put differently, the rule will only fail to perform well if the opponent plays so as to keep the player near indifferent. Moreover, it is easy to see that no deterministic learning rule can be consistent in all games against all possible opponent's rules: For example, in matching pennies given any deterministic rule it is easy to construct an opposing rule that beats it in every period. This suggests that a possible fix would be to randomize when nearly indifferent, and indeed Fudenberg & Levine (1995) showed that SFP is universally consistent.

This universality property (called worst-case analysis in computer science) has proven important in the theory of learning, perhaps because it is fairly easy to achieve. But getting the frequencies asymptotically right is a weak criterion, as for example it allows a player to ignore the existence of simple cycles. Aoyagi (1996), studied an extension of fictitious play in which agents test the history for "patterns," which are sequences of outcomes. Agents first check for the pattern of length 1 corresponding to yesterday's outcome, and count how often this outcome has occurred in the past. Then they look at the pattern corresponding to the two previous outcomes, and see how often it has occurred, and so on. Player i "recognizes" a pattern p at history h if the number of its occurrences exceeds an exogenous threshold that is assumed to depend only on the length of p. If no pattern recognized, beliefs are the empirical distribution. If one or more patterns detected, pick one pattern (rule for picking which one can be arbitrary) and let beliefs be a convex combination of the empirical distribution and the empirical conditional distribution in periods following this pattern. He shows that this form of

---

[18] The Hofbauer&Sandholm (2002) result mentioned above showed that this same symmetry condition applies to smooth best responses generated by stochastic payoff shocks.

pattern detection has no impact on the long-run outcome of the system under some strong conditions on the game being played.

Lambson & Probst (2004) considered learning rules that are a special case of those in Aoyagi's paper, and derive a result for general games: if the two players use equal patterns lengths and exact FP converges, then empirical c.d.f. of play converges to the convex hull of the set of NE. We expect that detecting longer patterns is an advantage. Lambson and Probst do not have general theorems about this, but they have an interesting example: In matching pennies, there is a pair of rules where player 1 has pattern length 0, player 2 has pattern length 1, and player 2 always plays a BR to player 1's anticipated action. Note that this claim lets us choose the two rules together. So specify that player 2's prior is that 1 will play *T* following the first time *(H, T)* occurs and *H* following the first time *(T ,H)* occurs . Suppose also that if players are indifferent they play *H,* and that they start out expecting opponent to play H. Then the first period outcome is (*H,T)*; next period is (*T,H)*, third period is (H,T) (because 1 plays H when indifferent) and so on.[19]

In addition, the basic model of universal consistency can be extended to account for some conditional probabilities. This can be done by directly estimating conditional probabilities using a sieve as described in Fudenberg & Levine (1999) or by the method of "experts" used in computer science. This method, roughly speaking, takes a finite collection of different "experts" corresponding to different dynamic models of how the data is generated, and shows that asymptotically it is possible in the worst case to do as

[19] If we specified that player 2 plays H whenever there is no data for the relevant pattern (e.g. that the "prior" for this pattern is that 1 plays T) then player 2 only wins 2/3 of the time.

well as the best expert.[20] That is, within the class of dynamic models considered, there is no reason to do less well than the best.

Calibration

While universal consistency seems an attractive property for a learning rule, it is fairly weak. Foster & Vohra (1997) introduced learning rules that are derived from calibrated forecasts. Calibrated forecasts can be explained in the setting of weather forecasts: Suppose that a weather forecaster sometimes says there is a 25% chance of rain, sometimes a 50% chance, and sometimes a 75% chance. Then looking over all his past forecasts, if on all the days when he said 25% chance of rain it actually rained 25% of the time, when he said 50% it rained 50% of the time and when he said 75% it rained 75% of the time, we would say that he was well calibrated. As Dawid (1985) pointed out, no deterministic forecast rule is calibrated in all environments; but just as with the related concept of universal consistency, calibration can be achieved with randomization, as shown by Foster and Vohra (1998). Calibration seems a desirable property for a forecaster to have, and there is some evidence that weather forecasters are in fact reasonably well calibrated (Murphy & Winkler, 1977), but the extent to which experimental subjects are well calibrated about their answers to trivia questions (e.g. "what is the area of Nigeria?") is under dispute (see e.g. Gigerenzer et al. (1991)).

In a game or decision problem, the question corresponding to calibration is: on all the occasions where a player took a particular action, how good a response was it? Put

[20] This work was initiated by Vovk (1990); it and subsequent developments are summarized in Fudenberg & Levine (1998). There are also a few papers in computer science that analyze other models of cycle detection combined with exact as opposed to smooth fictitious play.

differently, choosing action A is a "prediction" that it is a best response. If we take the frequency of opponents' play on all those periods where that prediction was made, we can ask: "was A actually a best response in those periods?" Or as in Hart and Mas-Colell (2000) we can measure this by regret: How much loss has the player suffered in those periods by playing A rather than the actual best response to the frequency over those periods? If, regardless of opponents' play the player is asymptotically calibrated in the sense that the time average regret for each action goes to zero, we say that the player is universally calibrated. Foster & Vohra (1997) showed that there are learning procedures that have this property, and moreover that if all players follow such rules the time average of the frequency of play must converge to the set of correlated equilibria of the game.

Because the algorithm originally used by Foster and Vohra involved a complicated procedure of finding stochastic matrices and their eigenvectors, one might ask whether it is a good approximation to assume that players follow universally calibrated rules. The "universal" aspect of universal calibration makes it impossible to empirically verify without knowing the actual rules that players use, but it is conceptually easy to tell whether learning rules are calibrated along the path of play: If they are, the time average of joint play converges to the set of correlated equilibria. If to the contrary some player is not calibrated along the path of play, she might notice that the environment is negatively correlated with her play, which should lead her to second-guessing her planned actions. For example, if it never rains when the agent carries an umbrella, she might think along the following lines: "I was going to carry an umbrella, so that means it will be sunny, so I should not carry an umbrella after all." Just like failures

of stationarity, some forms of non-calibration are more subtle and difficult to detect, but even so universally calibrated learning rules need not be exceptionally complex.

We will not focus on the algorithm of Foster and Vohra. Subsequent research has greatly expanded the set of rules known to be universally calibrated, and greatly simplified the algorithms and methods of proof. In particular, universally consistent learning rules may be used to construct universally calibrated learning rules by solving a fixed-point problem, which roughly corresponds to solving the fixed point problem of second guessing whether to carry an umbrella. This fixed point problem is a linear problem that is solved by inverting a matrix, as shown in Fudenberg & Levine (1998); the bootstrapping approach was subsequently generalized by Hart & Mas-Colell (2000).

Although inverting a matrix is conceptually simple, one may still wonder whether it is simple enough for people to do in practice. Consider the related problem of arbitrage pricing, which also involves inverting a matrix. Obviously people tried to arbitrage before there were computers or simple matrix inversion routines. Whatever method they used seems to have worked reasonably well, because examination of price data does not reveal large arbitrage opportunities (see e.g. Black & Scholes (1971) and Moore & Juh (2006).) That actual matrix inversion works better may be seen by the fact that large Wall Street arbitrage firms do not invert matrices by the seat of their pants, but by explicit calculations on a computer.[21]

We should also point out the subtle distinction between being calibrated and universally calibrated. For example, Hart & Mas-Colell (2001) examined simple algorithms that lead to calibrated learning, even though they are not universally

calibrated. Formally the fixed point problem that needs to be solved for universal calibration has the form $Rq = R^T q$ where q are the probabilities of choosing different strategies, and R is a matrix in which each row is the probability over actions derived by applying a universally consistent procedures to each conditional history of a players own play. Suppose in fact the player played action a last period. Let μ be a large number, and consider then defining current probabilities by

$$q(b) = \begin{cases} (1/\mu)R(a,b) \\ 1 - \sum_{c \neq a} q(c) \end{cases} \begin{vmatrix} b \neq a \\ b = a \end{vmatrix}.$$

Although this rule is not universally calibrated, Hart & Mas-Colell (2000) showed that it is calibrated provided everyone else uses similar rules. Cahn (2001) showed that the rule is also calibrated provided that everyone else uses rules that change actions at a similar rate. Intuitively, if other players do not change their play very quickly the procedure above implicitly inverts the matrix needed to solve $Rq = R^T q$.

Testing

One interpretation of calibration is that the "learner" has passed a test for learning, namely getting the frequencies right asymptotically, even though the "learner" started by knowing nothing. This has led to a literature that asks when and whether a person ignorant of the true law generating signals could fool a tester. Sandroni (2003) proposed two properties for a test: it should declare pass/fail after a finite number of periods, and it should pass the truth with high probability. If there is an algorithm that can

[21] The ability to implement sophisticated algorithms, both on modern computers and within the brain, shows the limitations of conclusions based on fMRI studies of simple learning tasks.

pass the test with high probability without knowing the truth, Sandroni says that it *ignorantly* passes the test. Sandroni showed that for any set of tests that give an answer in finite time and pass the truth with high probability, there is an algorithm that can ignorantly pass the test.

Subsequent work has shown some limitations of this result. Dekel & Feinberg (2006) and Olszeiwski & Sandroni (2006) relaxed the condition that the test yield a definite result in finite time. They showed that such a test can screen out ignorant algorithms, but only by using counter-factual information. Fortnow & Vohra (2008) showed that an ignorant algorithm that passes certain tests must necessarily be computationally complex, and Al-Najjar & Weinstein (2007) who showed that it is much easier to distinguish which of two learners is informed than to evaluate one learner in isolation. Feinberg & Stewart (2007) consider the possibility of comparing many different experts, some real and some false, and show that only the true experts are guaranteed to pass the test no matter what the other experts do.

Convergence to Nash Equilibrium

There are two reasons we are interested in convergence to Nash equilibrium. First, Nash equilibrium is widely used in game theory, so it is important to know when learning rules do and do not lead to Nash equilibrium. Second, Nash equilibrium (in a strategic form game) can be viewed as characterizing situations where no further learning is possible; conversely when learning rules do not converge to Nash equilibrium some agent could gain by using a more sophisticated rule.

This question has been examined by examining a class of learning rules to determine whether Nash equilibrium is reached when all players employ learning rules in

the class. For example, we have seen that if all players employ universally calibrated learning rules, then play converges to the set of correlated equilibrium. But this means that players may be correlating their play through the use of time as a correlating device, and why should players not learn this? In particular, are there classes of learning rules that when employed by all players lead to global convergence to a Nash equilibrium?

Preliminary results in this direction were negative. Learning rules are called "uncoupled" if the equation of motion for each player does not depend on the payoff function of the other players. (It can of course depend on their actual play.) Hart & Mas-Colell (2003) showed that uncoupled and stationary deterministic continuous-time adjustment systems cannot be guaranteed to converge to equilibrium in a game; this result has the flavor of Saari & Simon's (1978) result that price dynamics that are uncoupled across markets cannot converge to Walrasian equilibrium. Hart & Mas-Colell (2006) proved that convergence to equilibrium cannot be guaranteed in stochastic discrete-time adjustment procedures in the "1-recall" case where the state the system is the most recent profile of play. They also refine the past convergence result of Foster & Young by showing that convergence can be guaranteed in stochastic discrete-time systems where the state corresponds to play in the preceding two periods.

Despite this negative result, there may be uncoupled stochastic rules that converge probabilistically to Nash equilibrium, as shown in a pair of papers by Foster and Young Their first (2003) paper on this topic showed the possibility of convergence with uncoupled rules. However, the behavior prescribed by the rules strikes us as artificial and poorly motivated. However, their (2006) obtained the same result with rules that are more plausible. Further results can be found in Young (2008).

In the Foster & Young stochastic learning model the learning procedure follows a "status quo" action which it re-evaluates periodically. These re-evaluations take place at infrequent random times. During the evaluation period, some other action, randomly chosen with probability uniformly bounded away from zero, is employed instead of the status quo action. That the times of re-evaluation are random assures a fair comparison between the payoffs of the two actions. If the status quo is "satisfactory" in the sense that the alternate action does not do too much better, it is continued on the same basis (being reevaluated again). If it fails then the learner concludes that the status quo action was probably not a very good action. However, rather than adopting the alternative action, the learner goes back to the drawing board and picks a new status quo action at random.

We have already seen that if we drop the requirement of convergence in all environments, sensible procedures such as fictitious play converge in many interesting environments, for example in potential games. A useful counterpoint is the Shapley counterexample discussed earlier, in which stochastic fictitious play fails to converge but instead approaches a limit cycle. Along this cycle, players act as if the environment is constant, failing to anticipate the fact that their opponent's play is changing. This raises the possibility of a more sophisticated learning rule in which players attempt to forecast each other's future moves. This type of model was first studied in Levine (1991), who showed that players who were not myopic, but somewhat patient, would move away from Nash equilibrium as they recognized the commitment value of their actions. Dynamics in the purely myopic setting of attempting to forecast the opponent's next play, is studied in Shamma & Arslan (2005).

To motivate the Shamma & Arslan model, consider the environment of smooth fictitious play with exponential weighting of past observations,[22] which has the convenient property of being time homogeneous, and limit attention to the case of two players. Let λ be the exponential weight and let $z_i(t)$ be the vector over actions of player i that takes on the value 1 for the action taken in period t and 0 otherwise. Then the empirical weight frequency of player $i's$ play is $\sigma_i(t) = (1-\lambda)z_i(t-1) + \lambda\sigma_i(t-1)$. In SFP, at time t player i plays a smoothed best response $\beta_i(\sigma_{-i}(t-1))$ to this empirical frequency. However, $\sigma_{-i}(t-1)$ measures what $-i$ did in the past, not what she is doing right now. So it is natural to think of extrapolating $-i$'s past play to get a better estimate of her current play. Shamma & Arslan, motivated by the use of "proportional derivative control" to obtain control functions with better stability properties, introduce an auxiliary variable r_{-i} with which to do this extrapolation. This auxiliary variable tracks σ_{-i}, so changes in the auxiliary variable can be used to forecast changes in σ_{-i}. Specifically, suppose that $r_{-i}(t) = r_{-i}(t-1) + \lambda(\sigma_{-i}(t-1) - r_{-i}(t-1))$, that is r_{-i} adjusts to reduce the distance to σ_{-i}. The extrapolation procedure is then to forecast player $-i$'s play as $\sigma_{-i}(t-1) + \gamma(r_{-i}(t) - r_{-i}(t-1))$, the case $\gamma = 0$ corresponding to the exponentially weighted FP, $\gamma > 0$ corresponds to giving some weight to the estimate $r(t)$ of the derivative. This estimate is essentially the most recent increment in σ_{-i} when λ is very large; smaller values of λ correspond to smoothing by considering past increments as well.

---

[22] They use a more complicated derivation from ordinary fictitious play.

Motivated by stochastic approximation, (which requires the exponential weighting in beliefs be close to 1) Shamma and Arslan then propose to study the continuous time analog of this system. For the evolution of the state variable $\dot{\sigma}_i = \phi[\,\beta_i(\sigma_{-i} + \gamma\dot{r}_{-i}) - \sigma_i\,]$, which comes from taking the expected value of the adjustment equation for the weighted average, where ϕ is the exponential weight.[23] The equation of motion for the auxiliary variable is just $\dot{r}_{-i} = \lambda(\sigma_{-i} - r_{-i})$.

From the auxiliary equation $\ddot{r}_{-i} = \lambda(\dot{\sigma}_{-i} - \dot{r}_{-i})$ so that $\dot{r}_{-i}$ will be a good estimate of $\dot{\sigma}_{-i}$ if $\ddot{r}_{-i}$ is small. When $\ddot{r}_{-i}$ is small Shamma & Arslan show that the system globally converges to a Nash equilibrium distribution. However, there are not good known conditions on fundamentals that guarantee this result. What is true, that in some cases of practical interest, such as the Shapley example, simulations show that the system does converge.

2C. Reinforcement Learning , Aspirations, and Imitation

Now we consider the non-equilibrium dynamics of various forms of boundedly-rational learning, starting with models in which players act as if they do not know the payoff matrix,[24] and do not observe (or do not respond to) opponent's actions. We then go on to models that assume players do respond to data such as the relative frequency and payoffs of the strategies that are currently in use.

---

[23] It should be noted that Shamma & Arslan (2005) choose the units of time so that $\phi = 1$. In these time units, it takes one unit of time to reach the best response, so that choosing $\gamma = 1$ means that the extrapolation attempts to "guess" what other players will be doing at the time full adjustment to the best response takes place. Shamma and Arslan give a special interpretation to this case, which they refer to as "system inversion."

[24] This behavior might arise either because players do not have this information or because they ignore it due to cognitive limitations However, there is evidence that providing information on opponents' actions

Reinforcement learning has a long history in the psychology literature. Perhaps the simplest model of reinforcement learning is the cumulative proportional reinforcement" or "CPR" studied by Laslier et al. (2001). In this process, utilities are normalized to be positive, and the agent starts out with initial weights $CU_k(1)$ to each action *k*. Thereafter, the process updates the score (also called a propensity) of the action that was played by its realized payoff, and does not update the scores of other actions. The probability of action *k* at time t is then $CU_k(t)/\sum_j CU_j(t)$. Note that the "step size" of this process – the amount that the score is updated – is stochastic, and depends on the history to date, in contrast to the $1/t$ increment in beliefs for a Bayesian learner in a stationary environment. [25]

With this rule, every action is played infinitely often: The cumulative score of action k is at least its initial value, and the sum of the cumulative payoffs at time t is at most the initial sum plus t times the maximum payoff. Thus the probability of action k at time t is at least $a/(b+ct)$ for some positive constants a,b,c and so the probability of never playing k after time *t* is bounded by the product $\prod_{\tau=t}^{\infty}(1-a/(b+c\tau))=0$.

To analyze the process further, Laslier et al. used results on "urn models": the state space is the number of balls of each type, and each period 1 ball is added, so that the step size is $1/t$. Here the balls correspond to the possible action profiles (a^1,a^2), and the state space has dimension $\#A^1 \bullet \#A^2$ equal to the number of distinct action profiles.

---

does change the way players adapt their play, see Weber (2003). Börgers et al. (2004) characterizes "monotone" learning rules for settings when players observe only their own payoff.

[25] Erev & Roth (1999) study a perturbed version of this model where every action receives a small positive reinforcement in every period; Hopkins (2002) studies a normalized version of the Er'ev and Roth process where the step size is deterministic and of order 1/*t*.

The point is that the number of occurrences of a given joint outcome can increase by at most rate $1/t$, and the number of times that each outcome has occurred is a sufficient statistic for the realized payoffs and the associated cumulative utility. One can then use stochastic approximation techniques to derive the associated ODE $\dot{x} = -x + r(x)$, where x is the fraction of occurrences of each type, and *r* is the probability of each profile as a function of the current state.

Laslier et al. showed that when "player 2" is an exogenous fixed distribution played by Nature, the ODE converges to the set of maximizing actions from any interior point, and moreover that the stochastic discrete-time CPR model does the same thing. Intuitively, the fact that the system cannot lock on to the wrong action comes from the facts that every action is played infinitely often (so that players can learn the value of each action) and that the step size converges to 0. Laslier et al. also analyzed systems with two agents, each using CPR (and so acting as if they were facing a sequence of randomly drawn opponents). Some of their proofs were based on incorrect applications of results on stochastic approximation due to problems on the boundary of the simplex; Beggs (2005) and Hopkins & Posch (2005) provided the necessary additional arguments, showing that even in the case of boundary rest points, reinforcement learning does not converge to equilibria that are unstable under the replicator dynamic, and in particular cannot converge to non-Nash states. Beggs showed that the reinforcement model converges to equilibrium in constant-sum 2x2 games with a unique equilibrium, and Hopkins & Posch showed convergence to a pure equilibrium in rescaled partnership games. Since generically every 2x2 game is either a rescaled partnership game or a rescaled constant-sum game, what these results leave open is the question of convergence

in games that are rescaled constant sum but not constant sum without the rescaling; work in progress by Hofbauer establishes that reinforcement learning does converge in all 2x2 games.

Hopkins (2002) studied several "perturbed" versions of CPR with slightly modified updating rules; in one version the update rule is the same as CPR except that the each period the score of every action is updated by an additional small amount λ. Using stochastic approximation, he related the local stability properties of this process to that of a perturbed replicator dynamic. He showed (roughly speaking) that if a completely mixed equilibrium is locally stable for all smooth best response dynamics, it is locally stable for the perturbed replicator, and that if an equilibrium is unstable for all smooth best response dynamics, it is unstable for the perturbed replicator.[26] He also obtained a global convergence result for a "normalized" version of perturbed CPR where the step size per period is $1/t$ independent of the history.

Börgers & Sarin (1997) analyzed a related (unperturbed) reinforcement model, where amount of reinforcement does not slow down over time but is instead a fraction γ, so that in a steady state environment the cumulative utility of every action that is played infinitely often converges to its expected value. Because the system does not slow down over time, the fact that each action is played infinitely often does not imply that the agent learns the right choice in a stationary environment, and indeed the system has positive probability of converging to a state where the wrong choice is made in every period. At a technical level, stochastic approximation results for systems with decreasing steps do not apply to systems with a constant step size. Instead, Börgers and Sarin looked at the limit

of the process as the adjustment speed γ goes to 0, and show that over finite time horizons the trajectories of the process converge to that of its mean field, which is the replicator dynamic. (The asymptotics are however different: for example in matching pennies the reinforcement model will eventually be absorbed at a pure strategy profile, while the replicator dynamic will not.) Börgers and Sarin (2000) extended this model to allow the amount of reinforcement to depend on the agent's "aspiration level." In some cases, the system does better with an aspiration level than in the base Börgers & Sarin (1997) model, but aspiration levels can also lead to suboptimal "probability-matching" outcomes.[27]

It is worth mentioning that an inability to observe opponent's actions does not make it impossible to implement SFP, or related methods, such as universally calibrated algorithms. In particular, in SFP what matters is the utility of different alternatives. For example, in the exponential case

$$\overline{BR}^i(\sigma^{-i})(s^i) = \frac{\exp(\beta u(s^i, \sigma^{-i})}{\sum_{s'^i} \exp(\beta u(s'^i, \sigma^{-i})}$$

it is not important that the player observe σ^{-i}, he merely needs to see $u(s^i, \sigma^{-i})$, and there are a variety of ways to use historical data on the player's own payoffs to infer this.[28] Moreover we conjecture that the asymptotic behavior of a system where agents

---

[26] As mentioned in section 2A, two "small perturbations" of the same best response function can have differing implications for local stability.

[27] At one time psychologists believed that probability matching was a good description of human behavior, but subsequent research showed that behavior moves away from probability matching if agents are offered monetary rewards or simply given enough repetitions of the choice. (Lee, 1971).

[28] See Fudenberg & Levine (1998) and Hart & Mas-Colell (2001).

learn in this way will be the same as with SFP, though the relative probabilities of the various attractors may change, and the speed of convergence will be slower.

Reinforcement learning requires only that the agent observe his own realized payoffs. Several papers suppose that agents can access the actions and perhaps the payoffs of other members of the population, and thus can imitate the actions of those they observe. Björnerstedt & Weibull (1996) studied a deterministic, continuum-population model, where agents receive noisy statistical information about the payoff of other strategies, and switch to the strategy that appears to be doing the best. Binmore & Samuelson (1997) studied a model of imitation with fixed aspirations in a large finite population playing a 2x2 game.[29] In the unperturbed version of the model, each period, one agent receives a "learn draw" and compares the payoff of his current strategy against the sum of a fixed aspiration level and an i.i.d. noise term. (The agent plays an infinite number of rounds between each learn draw so that this payoff corresponds to the strategy's current expected value.) If the payoff is above the target level the agent sticks with his current strategy, otherwise he imitates a randomly chosen individual. In the perturbed process, the agent "mutates" to the other strategy with some fixed small probability λ. Binmore & Samuelson characterized the iterated limit of the invariant distribution of the perturbed process as first the population size goes to infinity and then the mutation rate shrinks to 0. In a coordination game this limit will always select one of the two pure-strategy equilibria, but the risk dominant equilibrium need not be selected,

[29] Fudenberg and Imhof (2008) generalize their assumptions and extend the analysis to games with an arbitrary finite number of actions.

because the selection procedure reflects not only the size of the "basins of attraction" of the two equilibria, but also the strength of the learning flow. [30]

A similar finding arises in the study of the frequency-dependent Moran process (Nowak et al., 2004) which represents a sort of imitation of successful strategies combined with the imitation of popular ones: When an agent changes his strategy, he picks a new one based on the product of the strategy's current payoff and its share of the population, so that if all strategies have the same current payoff, the probabilities of adoption exactly equal the population shares, while if one strategy has a much higher payoff, its probability of being chosen can be close to one. In the absence of mutations ot other perturbations, the Binmore & Samuelson (1997) and the Nowak et al. (2004) models both have the property that every "homogeneous" state where all agents play the same strategy is absorbing, while every state where two or more strategies are played is transient. Fudenberg & Imhof (2006) gave a general algorithm for computing the limit invariant distribution in these sorts of models for a fixed population size as the perturbation goes to 0, and applied it to 3x3 coordination games and to the model of Nowak et al. Benaïm & Weibull (2003) provided mean field results for the large-population limit of a more general class of systems, where the state corresponds to a mixed strategy profile, only one agent changes play per period, and the period length goes to 0 as the population goes to infinity.

Karandikar et al. (1998), Posch & Sigmund (1999), and Cho & Matsui (2004) analyzed endogenous aspirations and inertia in two-action games. In their models, a fixed pair of agents play each other repeatedly; the agents tend to play the action they played in

[30] See Fudenberg and Harris (1992) for a discussion of the relative importance of the size of the basin and

the previous period unless their realized payoff is less than their aspiration level, where the aspiration level is the average of the agent's realized payoffs.[31] In Posch and Sigmund, the behavior rule is simple: if the payoff is at least the aspiration level, then play the same action with probability $1-\varepsilon$ and switch; symmetrically, if the realized payoff is less than the aspiration level then switch with probability $1-\varepsilon$. In Krandikar et al, the agent never switches if his realized payoff is at least the aspiration level; as the payoff drops below the aspiration the probability of switching falls continuously to a lower bound *p*. Cho & Matsui also considered a smooth increasing switching function; in contrast to Krandikar et al. is they assume that there is a strictly positive probability of switching if the payoff is in a neighborhood of the current aspiration level.

The key aspect of these models is that because the aspirations update at rate $1/t$, they eventually move much more slowly than behavior. This allows Cho & Matsui to apply stochastic approximation techniques and relate the asymptotic behavior of the system to that of the system $\dot{a} = u_a - a$, where a is the vector of aspiration levels, and u_a is the vector of average payoffs induced by the current aspiration level. (This vector is unique because each given aspiration level corresponds to an irreducible Markov matrix on actions.[32]) Cho & Matsui concluded that their model leads to coordination on the Pareto-efficient equilibrium in a symmetric coordination game, and that play can converge to "always cooperate" in the prisoner's dilemma, provided that the gain from cheating is sufficiently small compared to the loss incurred when the other player cheats.

---

the cost of "swimming against the flow."

[31] Posch & Sigmund also analyze a variant where the aspiration level is yesterdays's payoff, as oppsed to the long-run average. Karandikar et al. focus on an extension of the model where aspirations are updated with noise.

Krandikar et al. obtained a similar result, except that it holds regardless of the gain to cheating: the difference comes from the fact that in their model agents who are satisfied stick with their current action with probability 1.

In these models, players do not explicitly take into account the fact that they are in a repeated interaction, but cooperation nonetheless occurs.[33] It is at least as interesting to model repeated interactions when players explicitly respond to their opponent's play, but the strategy space in a repeated game is large, so analyzes of learning dynamics have typically either restricted attention to a small subset of the possible repeated game strategies or analyzed related games where the strategy space is in fact small. The first approach has a long tradition in evolutionary biology, going back to the work of Axelrod & Hamilton (1981). Nowak et al. (2004) and Imhof et al. (2005) adopted it in their applications of the Moran process to the repeated prisoner's dilemma: The first paper considers only the two strategies "Always Defect" and "Tit for Tat", and shows that Tit for Tat is selected, essentially because its basin becomes vanishingly small when the game is played a large number of rounds. The second paper adds in the strategy "always C," which is assumed to have a small complexity-cost advantage over Tit for Tat; the result is cycles that spend most of the time near "All Tit for Tat" if the population and the number of rounds are large.[34] Jehiel (1999) considers a different sort of simplification: he supposes that players only care about payoffs for the next *k* periods, and believe that their opponent's play only depends on the outcomes in the past *m* periods.

---

[32] In Posch and Sigmund, behavior is not a continuous function of the state, but they use simulations to support the use of a similar equation.

[33] It is often possible to do well by being less than fully rational. This is especially important where precommitment is an issue: here it is advantageous for opponents to think you are irrationally committed. An interesting example of such a learning rule and a typical result can be found in Acemoglu & Yildiz (2001).

Instead of imposing restrictions on the strategy space or beliefs, one can consider an overlapping generations framework where players play just once, as in the "gift-giving game," where young people may give a gift to an old person. Payoffs are such that it is preferable to give a gift when young and receive one when old then to neither give nor receive a gift. This type of setting was originally studied without learning by Kandori (1992) who allowed "information systems" to explicitly carry signals about past play, and proved a folk theorem for a more general class of overlapping-generations games. Johnson, Pesendorfer & Levine (2001) showed that a simple red/green two signal information system can be used to sustain cooperation and that this emerges as the limit of the invariant distribution under the myopic best response dynamic with mutations. Nowak & Sigmund (1998a, b) offered an interpretation of Kandori's information systems as a public image, and use simulations of a discrete-time replicator process to argue that play converges to a cooperative outcome.

Pesendorfer & Levine (2007) studied equilibrium selection a related game under the "relative best reply dynamic," which says that players select best reply to the current state among the strategies that are currently active. To make the process ergodic, they assume that there are small perturbations corresponding both to imitation (copy a randomly chosen agent) and mutation, with imitation much more likely than mutation, Pesendorder and Levine then analyzed the limiting invariant distribution in games in which player simultaneously receive signals of each other's "intentions" and use strategies that simultaneously indicate intention and respond to signals about the other player's intention. These games always have trivial equilibria in which the signals are

---

[34] The result requires that the number of rounds is large given the population size.

ignored. Depending on how strong the signal is, there can be more cooperative equilibria. For example if players receive a perfect indication of whether their opponent is using the same strategy as they are, then the strategy of maximizing joint utility when the opponent is the same, but minmaxing the difference in utilities when the opponent is different is an equilibrium. Moreover, Pesendorfer and Levine showed that this equilibrium is selected n the limit of small perturbations.

3. Learning in Extensive-Form Games

3A. Information and Experimentation

In many settings with simultaneous moves, it seems natural for each player to observe the strategies used by each of his opponents after each play of the game. In extensive-form games, it seems more natural to assume that players observe at most which terminal nodes are reached, so that they do not observe how their opponents would have played at information sets that were not reached. To begin, we will briefly review the earliest work on this topic, which is based on the idea that if a player never plays a specific action, he may never observe how his opponents react to it, so incorrect beliefs about off-path play could persist, and play might converge to a non-Nash outcome.

More precisely, incorrect beliefs about off-path play can persist unless for some reason players obtain "enough" observations of off-path play. This raises three questions: 1) What outcomes can persist if there are very few observations of off-path play? 2) How much of off-path play is needed to imply that any long-run outcome satisfy the conditions of standard equilibrium conditions such as Nash equilibrium and sequential equilibrium? 3) How much off-path play will in fact occur under various models of learning?

The answer to what types of outcomes can persist in the absence of information about off-path play is given by the notion of self-confirming equilibrium (SCE). There are several versions of SCE. The most straightforward to define is that of unitary SCE. This requires that each player have beliefs μ_i over opponents play (ordinarily the space of their behavior strategies) that satisfies two basic criteria. First, players should optimize relative to their beliefs. Second, beliefs should be correct at those information sets on the game tree that are reached with positive probability. Put differently, the beliefs must assign probability one to the set of opponent behavior strategies that are consistent with actual play at those information sets. Even this version of SCE allows outcomes that are not Nash equilibria, as shown by an example of Fudenberg & Kreps (1988), but it is outcome-equivalent to Nash equilibrium in 2 player games (Battigalli (1987), Fudenberg & Kreps, (1995)).[35] One important variation on this basic definition, is the concept of heterogeneous SCE, which applies when there is a population of agents in each player role, so that different agents in the same player role can have different beliefs, but the beliefs of each agent must be consistent with what the agent observes given its own choice of pure strategy.

Although even unitary SCE is less restrictive than Nash equilibrium, it is by no means vacuous. For example, Fudenberg & Levine (2005) showed that self-confirming equilibrium is enough for the no-trade theorem. Basically, if players make a purely speculative trade, some of them have to lose, and they will notice this.

---

[35] More generally, unitary SCE with independent beliefs are outcome-equivalent to Nash equilibria in games with observed deviators. Kamada (2008) fixes an error in the original Fudenberg and Levine (1993a) proof of this, which relied in the claim that that "consistent" unitary, independent SCE were outcome-equivalent to Nash equilibria. The definition given of consistency was too weak for this to be true, Kamada give the appropriate definition.

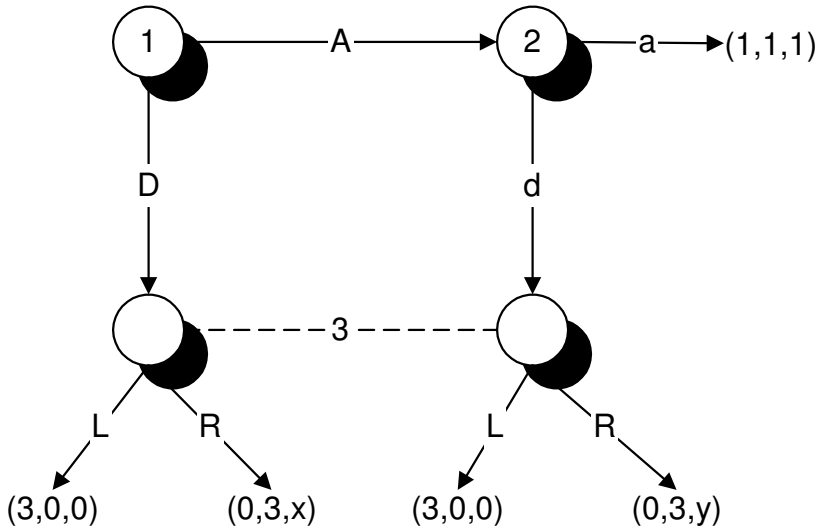
We turn now to the question of when there is enough experimentation to lead to a a stronger notion of equilibrium than SCE. Fudenberg & Kreps (1994) showed that non-Nash outcomes cannot persist if at every action is played infinitely often at every information on the path of play, and observe that refinements such as sequential equilibrium require in addition that every action is played infinitely often at other information sets as well. If behavior rules satisfy their "MME" condition, then actions are indeed played infinitely often on the path of play, and moreover every action is played infinitely often in games of perfect information; for this reason the only stable outcomes in such games are the backwards induction solutions. However, they point out that the MME condition requires more experimentation than may be plausible, as it is not clear that players at seldomly-reached information sets will choose to do that much experimentation. This is related to the fact that if each player experiments at rate 1/t in period t, then players on the path of play experiment infinitely often (because $\sum_{t=1}^{\infty} 1/t = \infty$) while players who are only reached when others experiment will experiment a finite number of time (because $\sum_{t=1}^{\infty} 1/t^2 \neq \infty$.) Fudenberg and Levine (1993b) examine endogenous experimentation. They derive experimentation rates from the play of patient expected-utility maximizers, and show that there is enough experimentation to rule out non-Nash outcomes when the discount factor is close enough to 1, but they do not address the question of whether players will experiment enough to rule out outcomes that are not subgame perfect.

Noldeke & Samuelson (1993) considered a large-population model of learning where "experiments" occur when the players "mutate" and change their beliefs and thus

their actions. In most periods players do not update their beliefs at all, but with some fixed probability a player receives a "learn draw," observes the terminal nodes in all matches, and changes his beliefs about play at all reached information sets to match the frequencies in his observation. In games of perfect information, this leads to a refinement of SCE, and in some special cases it leads to subgame perfection.

Dubey & Haimanko (2004) studied a similar model of learning in a game of perfect information, where agents play best responses to their beliefs, and beliefs are updated to be consistent with observed data; the model is quite flexible as to what this means, as it allows players to consider only the most recent observation or older data as well. They show that the system converges to a (unitary) self-confirming equilibrium with independent beliefs; because this is a game with identified deviators, the steady state is thus outcome-equivalent to a Nash equilibrium.

The belief-based models mentioned above place no constraints on the players' beliefs other than consistency with observed data, and in particular are agnostic about what prior information any player might have about the payoff functions of the others. Rubinstein & Wolinksy (1994) and Dekel et al. (1999) added the restriction that players do know the payoff functions of the others, leading to the concepts of 'rationalizable conjectural equilibrium" and "rationalizable self-confirming equilibrium." To see the difference that this makes, consider Dekel et al's variation on the game Fudenberg and Kreps used to show the non-equivalence of SCE and unitary SCE with independent beliefs:



Here if x and y have the same sign, then player 3 has a conditionally dominant strategy, and so if players 1 and 2 assign high probability to a neighborhood of 3's true payoffs then they must have very similar beliefs about his play. In this case (A,a) is not the outcome of any RSCE. However, if x and y have opposite signs, then even common knowledge of 3's payoff function does not imply common beliefs about his play, and RSCE allows (A,a).[36] Another variation on this theme is the SCE-motivated robustness notion of Espinoza (2008). He allowed players to be uncertain about hierarchies of beliefs, but these hierarchies of beliefs are required to be consistent with players' knowledge of equilibrium play. Rationalizability, Nash equilibrium and SCE are special cases. Despite the broad range of possibilities allowed, the solution may be characterized by an iterative elimination procedure.

Lehrer & Solan's (2007) "partially specified equilibrium" is a variant of SCE where players observe a partition of the terminal nodes. A leading example is the trivial

[36] Dekel et al. only defined RSCE for the unitary case; the appropriate heterogeneous definition, and its consequences, is still unresolved.

partition which provides no information at all. While this on its own would allow a great multiplicity of beliefs (and only rule out the play of dominated strategies) the solution concept pins down beliefs by the worst-case assumption that players maximize their expected payoff against the confirmed belief that gives the lowest payoff. With this trivial partition, the unique PSE in a symmetric coordination game is for each player to randomize ½- ½ , which is not a SCE. At the other extreme, with the discrete partition on terminal nodes, the PSE must be a SCE.

3B. Solution Concepts and Steady-State Analysis

Self-confirming equilibrium is based on the idea that player should have correct beliefs about probability distributions that they observe sufficiently often, so that the specification of the "observation technology" is essential. The original definition of SCE assumes that players observe the terminal node that is reached, but in some settings it is natural to assume that they observe less than this. For example in a sealed-bid auction, players might only observe the winning bid and the identity of the winning bidder, but observe neither the losing bids nor the types of the other players. In the setting of a static Bayesian game, Dekel et al. (2004) extended the definition of SCE to allow for these sorts of coarser maps from outcomes of the game to observations. If players do observe the outcome of each round of play, meaning both the actions taken and the realization of Nature's move, the set of self-confirming equilibria is the same as the set of Nash equilibria with a common prior; Dekel et al. pointed out that the same conclusion applies if players observe the actions played and there are private values, so that each player's private information relates only to their own payoff. When players do not observe the actions played, or there are not private values, the set of self-confirming equilibria can

include non-Nash outcomes. Dekel et al. argued that Nash equilibrium without a common prior is difficult to justify as the long-run result of a learning process, because it takes very special assumptions for the set of such equilibria to coincide with the set of steady states that could arise from learning. Intuitively, Nash equilibrium requires that players have correct beliefs about the strategies their opponents use to map their types to their actions, and in order for repeated observations to lead players to learn the distribution of opponents' strategies, the signals observed at the end of each round of play must be sufficiently informative. Such information will tend to lead players to also have correct and hence identical beliefs about the distribution of Nature's moves.

While SCE assumes that the players' inferences are consistent with their observations, a related strand of this literature assumes that players make systematic mistakes in inference. The leading example here is the Jehiel (2005) notion of analogy-based expectations equilibrium or "ABEE," where players group the opponents' decision nodes into "analogy classes," with the player believing that play at each node in a given class is identical. Given this, the player's beliefs must then correspond to the actual average of play across the nodes in the analogy class.

This is most easily illustrated by an example. Consider a game of perfect information, where Nature moves first, choosing state A with probability 2/3 or state B with probability 1/3, player 1 moves second, choosing either action A1 or action B1, with player 2 moving last, again choosing either action A2 or action B2. We will suppose for illustrative purposes that player 2 is a dummy receiving a zero payoff no matter what, and that player 2 chooses A in state A and B in state B regardless of what player 1 did. Player 1 gets one if his action matches that of player 2 and zero if not. Then in state A player 1

should play A1 and in state B player 1 should play B1. However, if player 1 views all nodes of player 2 following a given move as belonging to an analogy class, then he believes that player 2 will play A2 2/3rds the time, regardless of the state, and so player 1 will play A1 regardless of the state. This is an example of an ABEE.

If player 1 observes and remembers the outcome of each game, then as he learns that player 2 plays A2 2/3rds of the time, he will also get evidence that player 2's play is correlated with the state. Thus if he is a rational Bayesian and assigns positive probability to player 2 observing the state, he should eventually learn that this is the case. Conversely, even a rational player 1 could maintain the belief that 2's play is independent of the state provided that he has a doctrinaire prior that assigns probability 1 to this independence. Such doctrinaire priors may seem unreasonable, but they are an approximation of circumstances where player 1 has a very strong prior conviction that player 2's play is independent of the state. In this case it will take a very long time to learn that this is not true. [37]

An alternative explanation for analogy-based reasoning is that players are boundedly rational, so that they are unable to remember all that they have observed, perhaps because at an earlier stage they chose not to expend the resources required for a better memory. In our example this would correspond to player 1 only being able to remember the fraction of time that 2 played A2 and not the correlation of this play with

[37] Ettinger & Jehiel (2005) explicitly recognized this issue, saying "From the learning perspective…it is important that a (player 1) does not play herself too often the game as the observation of past performance might trigger the belief that there is something wrong with (player 1)'s theory."

the state; this is analogous to SCE when the player 1's end-of-stage observation is simply player 2's action, and includes neither Nature's move nor player 1's realized payoff.[38]

Ettinger & Jehiel (2005) considered how a fully rational opponent might manipulate the misperceptions of an opponent who reasons by faulty analogy. They referred to this as "deception" and give a number of applications, as well as relating the idea to the "Fundamental Attribution Error" of social psychology. Jehiel & Koessler (2008) provided additional applications in the context of one-shot two-player games of incomplete information, and study in particular the conditions for successful coordination in a variety of games. They also study information transmission, and show that with analogy based reasoning, the no-trade theorem may fail, in contrast to the positive result under SCE. These many applications, while interesting, suggest that very little is ruled out by ABEE absent some constraints on the allowed analogy classes. Developing a taxonomy of ABEE's implications could be useful, but it seems more important to gain a sense of which sorts of false analogies are relevant for which applications and ideally to endogeneize the analogy classes.

ABEE is closely related to Eyster & Rabin's (2005) notion of cursed equilibrium. This focuses specifically on Bayesian games, and assumes "analogy" classes of the form that opponents' play is independent of their types. However, they introduce a "cursedness" parameter, and assume that each player's beliefs are a convex combination of the "analogy" based expectations and the correct expectations. When the cursedness parameter equals zero, we have the usual Bayesian equilibrium, when it is one we have in effect ABEE. Less obviously, by changing the information structure, it is possible to

[38] Since player 1 does observe Nature's move in the course of play, this form of SCE incorporates a form of

represent cursed equilibria as ABEE for intermediate parameter values as well. Miettinen (2007) shows how to find the correct information partition, and proves the equivalence. ABEE is also related to the "valuation equilibrium" of Jehiel & Samet (2005), where beliefs about continuation values take the place of beliefs about moves by opponents and Nature; the relationship between the outcomes allowed by these two solution concepts has not yet been determined.

In a related but different direction is the work of Esponda (2008). He supposed that there are two sorts of players: sophisticated players whose beliefs are self-confirming, and naïve players whose marginal beliefs about actions and payoff realizations are consistent with the data but who can have incorrect beliefs about the joint distribution. He then showed how in an adverse selection problem, the usual problem of self-selection is exacerbated. Interestingly, whether a bias can arise in equilibrium in this model is endogenous.

The work of Acemoglu et al. (2007) studied when rational Bayesian agents can maintain different beliefs even when faced with a common infinite data set. In their model, players are learning about a fixed unknown parameter, and update their beliefs using different likelihood functions. The agents' observations do not identify the underlying parameter, which is why the agents can maintain different beliefs about the parameter even though they asymptotically agree about the distribution of signals. Of course this lack of identification only matters if the unknown parameter is payoff-

---

imperfect recall.

relevant; loosely speaking, the assumptions in Acemolgu et al. correspond to the case in Dekel et al. (2004) where agents observe neither Nature's move nor their own payoffs.[39]

3C. Learning Backwards Induction

Now we turn to the question of when there will be enough experimentation to lead to restrictions beyond Nash equilibrium. As we discussed above, the earlier literature gave partial results in this direction. The more recent literature has focused on the special case of games of perfect information, and the question of when learning leads to the backwards induction outcome.

In a game of perfect information with generic payoffs, so that there are no ties, we should expect that many reasonable learning procedures will converge to subgame perfection provided that there is "enough" experimentation, and in particular if players experiment with a fixed non-vanishing probability. In this case, since all the final decision nodes are reached infinitely often, players will learn to optimize there; eventually players who move at the immediately preceding nodes will learn to optimize against the final-node play, and so forth. This backwards induction result, not surprisingly, is quite robust to the details of the process of learning. For example, Jehiel & Samet (2005) considered a setting where players use a valuation function to assess the relative merit of different actions at a node. Valuations are determined according to historical averages the moves have earned, so this without experimentation this is equivalent to fictitious play on the agent-normal form. When a small fixed amount of exogenous experimentation or "trembles" is imposed on the players, every information

[39] Acemoglu et al. (2007) suggest that the assumptions are a good description of learning whether Iraq had weapons of mass destruction.

set is reached infinitely often, so any steady state must approximate a Nash equilibrium of the agent-normal form and thus subgane-perfect; Jehiel and Samet showed moreover that that play does indeed converge, and provide some additional results about attaining individually rational payoffs in games more general than games of perfect information. Indeed, this is true in general games, regardless of how the other players play.[40]

In a different direction, Laslier & Walliser (2002) considered the "cumulative proportional reinforcement" learning rule. Here a player chooses a move with a probability proportional to the cumulative payoff she obtained in the past with that move. Again, when all player employ this learning rule, the backwards induction equilibrium always results in the long-run. Hart (2002) considered a model of myopic adjustment with mutations in a large population. Players are generally locked in to particular strategies, but are occasionally allowed to make changes. When they do so, they with very high probability choose to best-respond to the current population of players; and with low probability "mutate" to a randomly chosen strategy. One key assumption is that the game is played in what he calls the "gene normal form," which is closely related to the agent normal form: in the gene normal form, instead of a separate player at each information set, there is a separate population of players at each information set, so best responses and mutations are chosen independently across nodes. Hart showed that the unique invariant distribution of the Markov evolutionary process converges to placing all weight on the backward induction equilibrium in the limit as the mutation rate goes to

[40] In the special case of a "win/lose" player who gets a payoff of either zero or one, and who has a strategy that gives him one against any opponent strategy, Jehiel and Samet showed that there is a time after which the win-lose player always wins, even if the valuation is simply given by last period's payoff.

zero and the population size goes to infinity, provided that the expected number of mutations per period is bounded away from 0; Gorodeisky (2006) showed that this last condition is not necessary.

All of the papers with positive results assume, in effect, exogenously given experimentation. However, the incentives to experiment depend on how useful the results will be: if an opportunity to experiment arises infrequently, then there is little incentive to actually carry out the experiment. This has implications for backwards induction explored in Fudenberg & Levine (2006), who re-examined the steady-state model of Fudenberg & Levine (1993b) in subclass of games of perfect information where each player moves only once on any path of play. The key observation is that for some prior beliefs experimentation takes place only on the equilibrium path, so a relatively sharp characterization of the limit equilibrium path (the limit of the steady state paths as first the lifetimes go to infinity and then the discount factor goes to 1) is possible. A limit equilibrium path must be the path of a Nash equilibrium, but must satisfy also the property that one step off the equilibrium path, play follows a self-confirming equilibrium. In other words, wrong or "superstitious" beliefs can persist, provided that they are at least two steps off the equilibrium path, so that they follow deviations by two players. The reason is that the second player has little incentive to experiment since the first deviator deviates infrequently, so information generated by the second experiment has little value as the situation is not expected to recur for a long time.

3D. Non-equilibrium Learning in Macroeconomics

Learning, especially passive learning, has long played a role in macroeconomic theory. Lucas's (1976) original rationale for rational expectations theory was that it is

implausible to explain the business cycle by assuming that people repeatedly make the same mistakes. The Lucas critique, that individual behavior under one policy regime cannot be reasonably thought to remain unchanged when the regime changes, is closely connected to the idea of self-confirming equilibrium (Fudenberg & Levine (2007)). Indeed, in recent years, the idea of self-confirming equilibrium has had many applications in macroeconomics, so much so that it is the central topic of Sargent's 2008 AEA Presidential Address. As this address is an excellent survey of the area, we limit ourselves here to outlining the broad issues of related to learning that have arisen in macroeconomics.

One of the important uses of learning theory in macroeconomics is to use dynamic stability as a way to select between multiple rational expectations or self-confirming equilibria. Several learning dynamics have been studied, most notably the robust learning methods of Hansen and Sargent (2001). A good set of examples of equilibrium selection using learning dynamics can be found in Evans & Honkapohja (2003). Much of the area was pioneered by Marcet & Sargent (1998a,b), and recent contributions include Cho et al. (2002) and Sargent & Williams (2005) who examined the dynamics of escaping from "Nash" inflation.

The application of SCE to study the role of misperceptions in macroeconomics has also been important. Historically, the government's misperception of the natural rate hypothesis played a key role in the formulation of economic policy. This is discussed by Sargent (1999), Cogley & Sargent (2005) and Primiceri (2006) among others. The narrower problem of commodity money and the melting of coins has also been studied using the tools of self-confirming equilibrium by Sargent and Velde (2002).

Alesina & Angeletos (2005) used SCE to analyze the political economy of tax policy. They observe that if wealth is due to luck optimal insurance implies a confiscatory tax is efficient. On the other hand if wealth is due to effort transfers should be low to encourage effort. But even if wealth is due to effort, if taxes are confiscatory, effort does not generate wealth, only luck does so beliefs that only luck matters will be self-confirming. They then used the resulting multiplicity of SCE to reconcile cross-country correlation of perceptions about wealth formation and tax policy. In a similar vein, Giordani & Ruta (2008) show how incorrect but self-confirming expectations about the skills of immigrants can explain cross-country variation in immigration policy.

Acknowledgments: We are grateful to Sergiu Hart, Josef Hofbauer, Bill Sandholm, Satoru Takahashi, Yuichi Yamamoto, and Peyton Young for helpful comments, and to NSF grants SES-03-14713 and SES- 06-646816 for financial support.

LITERATURE CITED

Acemoglu D, Chernochozhukov V, Yildiz M. 2007. Learning and Disagreement in an Uncertain World, mimeo.

Acemoglu D, Yildiz M. 2001. Evolution of Perceptions and Play, mimeo

Alesina A, Angeletos G.-M. 2005 Fairness and Redistribution. *American Economic Review*, 95: 960-80.

Al-Najjar N, Weinstein J. 2007. Comparative Testing of Experts, mimeo.

Aoyagi M. 1996 Evolution of Beliefs and the Nash equilibrium of normal form games. *Journal of Economic Theory* 70:444-69.

Axelrod A, Hamilton W. 1981. The Evolution of Cooperation, *Science.* 211:1390-96.

Banos, A 1968. On Pseudo-Games. *Annals of Mathematical Statistics.* 39:1932-1945.

Battigalli P. 1987. *Comportomento razioanle ed eqiulibrio nei giochi e nelle situazioni sociali,* unpublished undergraduate dissertation, Bocconi University, reprinted in .translation in *Decisions, Games and Markets,* ed. Battigalli, Montesano, and Panunzi, Springer, 1997.

Beggs AW. 2005 On the Convergence of Reinforcement Learning. *Journal of Economic Theory* 122:1-36.

Benaïm M, Hirsch M. 1999. Mixed Equilibria Arising from Fictitious Play in Perturbed Games, *Games and Economic Behavior* 29: 36-72.

Benaïm M, Hofbauer J, Hopkins E. 2005. Learning in Games with Unstable Equilibria. mimeo.

Benaïm M, Hofbauer J, Sorin S. 2006 Stochastic Approximation and Differential Inclusions, Part II: Applications. *Mathematics of Operations Research* 31:673-695.

Benaim M Raimond O. 2007. Simulated Annealing, Vertex Reinforced Random Walks and Learning in Games, mimeo.

Benaïm M, Weibull J. 2003 Deterministic Approximation of Stochastic Evolution in Games. *Econometrica* 71: 873-903.

Binmore K, Samuelson L. 1997. Muddling Through: Noisy Equilibrium Selection. *Journal of Economic Theory* 74: 235-65.

Björnerstedt J, Weibull J. 1996. Nash Equilibrium and Evolution by Imitation," in *The Rational Foundation of Economic Behavior,* ed K. Arrow et al.London, MacMillan.

Black F, Scholes, M. 1972. The Valuation of Option Contracts and a Test of Market Efficiency, *The Journal of Finance* 27: 399-417

Blackwell, D. 1956a. An Analog of the Minimax Theorem for Vector Payoffs. *Pacific Journal of Mathematics* 6,1-8.

Blackwell D. 1956b. Controlled Random Walks, in *Proceedings of the International Congress of Mathematicians*, 3, North Holland Press, Amsterdam, pp. 336-338.

Börgers T, Morales A, Sarin R. 2004. Expedient and Monotone Learning Rules *Econometrica* 72: 383-405.

Börgers T, Sarin R. 1997. Learning Through Reinforcement and the Replicator Dyamics. *Journal of Economic Theory* 77:1-14.

Börgers T, Sarin R. 2000. Naïve Reinforcement Learning with Endogenous Aspirations. *International Economic Review* 41: 921-50.

Cahn, A. 2001. General Procedures Leading to Correlated Equilibria. *International Journal of Game Theory* 33: 21-40.

Camerer C, Ho Y. 1999. Experience-Weighted Attraction Learning in Normal Form Games. *Econometrica* 67: 837-94.

Cheung YW, Friedman D. 1997. Individual Learning in Normal Form Games: Some Laboratory Results. *Games and Economic Behavior* 19:46-76.

Cho I-K, Matsui A. 2005. Learning Aspiration in Repeated Games. *Journal of Economic Theory* 124:171-201.

Cho I-K, Sargent TJ. 2002. Escaping Nash Inflation. *Review of Economic Studies* **69**: 1-40.

Cogley T, Sargent TJ. 2005. The Conquest of U.S. Inflation:Learning and Robustness to Model Uncertainty. *Review of Economic Dynamics* 8: 528-63.

Dekel E, Feinberg Y. 2006. Non-Bayesian Testing of a Stochastic Prediction. *Review of Economic Studies* 73: 893-906.

Dekel E, Fudenberg D, Levine DK. 1999. Payoff Information and Self-Confirming Equilibrium. *Journal of Economic Theory* 89: 165-85.

Dekel E, Fudenberg D, Levine DK. 2004. Learning to Play Bayesian Games. *Games and Economic Behavior* 46: 282-303.

Dubey P, Haimanko O. 2004. Learning with Perfect Information. *Games and Economic Behavior* 46: 304-24.

Ellison G, Fudenberg D. 2000. Learning Purified Equilibria *Journal of Economic Theory* 90: 84-115.

Ely J, Sandholm WH. 2005. Evolution in Bayesian games I: Theory, *Games and Economic Behavior* 53: 83-109,

Erev I, Roth A. 1998. Predicting How People Play Games: Reinforcement Learning in Games with Unique Strategy Mixed-Strategy Equilibrium. *American Economic Review* 88: 848-881.

Ettinger D. Jehiel P. 2005. Towards a Theory of Deception, mimeo.

Esponda I. 2008. Behavioral Equilibrium in Economies with Adverse Selection *American Economic Review*. Accepted subject to minor revisions.

Evans R, Honkapohja S. 2003. Expectations and the Stability Problem for Optimal Monetary Policies. *Review of Economic Studies* 70: 807-24.

Eyster E, Rabin M. 2005. Cursed equilibrium. *Econometrica* 73: 1623-72.

Feinberg Y, Stewart C. Testing Multiple Forecasters. Mimeo Stanford Graduate School of Business.

Fortnow L, Vohra R. 2008. The Complexity of Forecast Testing. *Proceedings of the 9th ACM conference on Electronic Commerce.*

Foster D, Vohra R. 1997 Calibrated Learning and Correlated Equilibrium. *Games and Economic Behavior* 21: 40-55.

Foster D, Vohra R. 1998. Asymptotic Calibration. *Biometrika* 85: 379-90.

Foster D, Young P. 2003. Learning, Hypothesis Testing, and Nash Equilibrium. *Games and Economic Behavior* 45: 73-96.

Foster D, Young P. 2006.Regret testing: learning to play a Nash equilibrium without knowing you have an opponent. *Theoretical Economics,* 1: 341-367.

Freedman D 1965. On the Asymptotic Behavior of Bayes Estimates in the Discrete Case II. *Annals of Mathematical Statistics* 34: 1386-1403.

Fudenberg D, Harris C. 1992 Evolutionary Dynamics with Aggregate Shocks. *Journal of Economic Theory*, 57: 420-41.

Fudenberg D, Imhof L. 2006. Imitation Processes with Small Mutations. *Journal of Economic Theory* 131: 251-62.

Fudenberg D, Imhof L. 2008. Monotone Imitation Dynamics in Large Populations *Journal of Economic Theory* 140: 229-45.

Fudenberg D, Kreps D. 1988. A Theory of Learning, Experimentation, and Equilibrium in Games. Mimeo.

Fudenberg D, Kreps D. 1993. Learning Mixed Equilibria. *Games and Economic Behavior* 5: 320-67.

Fudenberg D, Kreps D. 1994 Learning in Extensive Games, II: Experimentation and Nash Equilibrium. mimeo.

Fudenberg D, Kreps D. 1995 Learning in Extensive Games, I: Self-Confirming Equilibrium. *Games and Economic Behavior* 8: 20-55.

Fudenberg D, Levine DK. 1993a. Self Confirming Equilibrium. *Econometrica* 61: 523-46.

Fudenberg D, Levine DK 1993b. Steady State Learning and Nash Equilibrium. *Econometrica* 61: 547-73.

Fudenberg D, Levine DK 1995. Consistency and Cautious Fictitious Play. The *Journal of Economic Dynamics and Control* 19: 1065-89.

Fudenberg D, Levine DK 1999a Conditional Universal Consistency *Games and Economic Behavior,* 29: 104-30.

Fudenberg D, Levine DK 1999b. An Easier Way to Calibrate. *Games and Economic Behavior* 29: 131-37.

Fudenberg D, Levine DK 2005. Learning and Belief-Based Trade. *The Latin American Journal of Economics* 42: 199-207.

Fudenberg D, Levine DK. 2006. Superstition and Rational Learning. *American Economic Review* 96: 630-51.

Fudenberg D, Levine DK. 2007. Self-Confirming Equilibrium and the Lucas Critique. Mimeo, forthcoming in the *Journal of Economic Theory.*

Fudenberg D, Takahashi S. 2007. Heterogeneous Beliefs and Local Information in Stochastic Fictitious Play, mimeo.

Gigerenzer G, Hoffrage U, Kleinbölting H. 1991. Probabilistic mental models: a Brunswikian theory of confidence, *Psychological Review* 98:506-28.

Giodanu P, Ruta M.2006. Prejudice and Immigration, mimeo.

Gorodeisky, Z. 2006. Evolutionary Stability for Large Populations and Backwards Induction. Mathematics of Operations Research 31:369-80.

Hannan J. 1957. Approximation to Bayes' risk in Repeated Play. In *Contributions to the Theory of Games, vol. 3*, ed. M. Dresher, A.W. Tucker, and P. Wolfe. Princeton, Princeton University Press, pp. 97-139.

Hansen L, Sargent T 2001 Robust Control and Model Uncertainty. *American Economic Review* 91:60-66.

Harsanyi J. 1973. Games with Randomly Disturbed Payoffs: A New Rationale for mixed-strategy equilibria. *International Journal of Game Theory* 2:1-23.

Hart S. 2002. Evolutionary Dynamics and Backward Induction. *Games and Economic Behavior* 41: 227-64.

Hart S, Mas-Colell A. 2000. A Simple Adaptive Procedure Leading to Correlated Equilibrium. *Econometrica* 68:1127-50.

Hart S, Mas-Colell A. 2001. A General Class of Adaptive Strategies. *Journal of Economic Theory* 98: 26-54.

Hart S, Mas-Colell A. 2003. Uncoupled Dynamics Do Not Lead to Nash Equilibrium *American Economic Review* 93: 1830-36.

Hart S, Mas-Colell A. 2006 Stochastic Uncoupled Dynamics and Nash Equilibrium. *Games and Economic Behavior* 57:286-303.

Hofbauer J, Sandholm W. 2002. On the Global Convergence of Stochastic Fictitious Play. *Econometrica* 70: 2265-94.

Hofbauer J, Sandholm W. 2007. Evolution in Games with Randomly Disturbed Payoffs. *Journal of Economic Theory* 132:47-69.

Hofbauer J, Sigmund K. 2003. Evolutionary Game Dynamics. *Bulletin of the American Mathematical Society* 40: 479-519.

Hopkins E, Posch M. 2005. Attainability of Boundary Points under Reinforcement Learning. *Games and Economic Behavior* 53: 110-25.

Hopkins E. 1999. Learning, Matching and Aggregation. *Games and Economic Behavior* 26: 79-110.

Hopkins E. 1999. A Note on Best Response Dynamics. *Games and Economic Behavior* 29:138-50.

Hopkins E. 2002. Two Competing Models of How People Learn in Games. *Econometrica* 70: 2141 -66.

Imhof I, Fudenberg D, Nowak M. 2005. Evolutionary Cycles of Cooperation and Defection. *Proceedings of the National Academy of Science* 102: 10797-800.

Jehiel P. 1999. Learning to Play Limited Forecast Equilibria. *Games and Economic Behavior* 22: 274-298.

Jehiel P. 2005. Analogy-based expectation equilibrium. *Journal of Economic Theory* 123: 81-104

Jehiel P, Koessler F. 2008. Revisiting Games of Incomplete Information with Analogy-based Expectations. *Games and Economic Behavior* 62:533-557.

Jehiel P, Samet D. 2005. Learning to Play Games in Extensive Form by Valuation. *Journal of Economic Theory* 124: 129-48.

Jehiel P, Samet D. 2007. Valuation Equilibrium. *Theoretical Economics* 2: 163-185.

Johnson P, Levine DK, Pesendorfer W. 2001.Evolution and Information in a Gift-Giving Game. *Journal of Economic Theory* 100: 1-22.

Jordan, J. 1995. Bayesian Learning in Repeated Games. *Games and Economic Behavior* 9: 8-20.

Kalai E, Lehrer E. 1993. Rational Learning Leads to Nash Equilibrium. *Econometrica* 61: 1019-45.

Kamada Y. 2008 Strongly Consistent Self-Confirming Equilibrium. mimeo.

Kandori M. 1992. Repeated Games Played by Overlapping Generations of Players. *Review of Economic* Studies 59: 81-92.

Karandikar K, Mookherjee D, Ray D, Vega-Redondo F. 1998. Evolving aspirations and cooperation. *Journal of Economic Theory* 80: 292–331.

Kreps D 1990 *Game Theory and Economic Modelling.* Oxford: Clarendon Press

Lambson V, Probst D. 2004. Learning by Matching Patterns. *Games and Economic Behavior* 46: 398-409.

Laslier JF, Topol R, Walliser B. 2001. A Behavioral Learning Process in Games. *Games and Economic Behavior* 37: 340-66.

Laslier JF, Walliser B. 2002. A Reinforcement Learning Process in Extensive Form Games. *International Journal of Game Theory* 33: 219-27.

Lee W. 1971 *Decision Theory and Human Behavior.* Wiley: New York

Levine DK. 1999. Learning in the Stock Flow Model. In *Money, Markets and Method: Essays in Honour of Robert W. Clower*, ed. P. Howitt, E. de Antoni and A. Leijonhufvud, Edward Elgar: Cheltenham, 236-246.

Levine DK, Pesendorfer W. 2007. Evolution of Cooperation through Imitation. *Games and Economic Behavior* 58: 293-315.

Lucas R. 1976. Econometric Policy Evaluation: A Critique. *Carnegie-Rochester Conference Series on Public Policy* **1**: 19.

Luce RD, Raiffa H. 1957. *Games and Decisions.* Wiley, New York.

Marcet A, Sargent TJ 1989a. Convergence of Least Squares Learning Mechanisms in Self Referential Linear Stochastic Models. *Journal of Economic Theory* 48: 337-68.

Marcet A, Sargent TJ 1989b. Convergence of Least-Squares Learning in Environments with Hidden State Variables and Private. *Journal of Political Economy* 97" 1306-22.

McKelvey P, Palfrey T. 1995 Quantal Response Equilibria for Normal Form Games. *Games and Economic Behavior* 10:6-38.

Megiddo N. 1980. On Repeated Games with Incomplete Information Played by Non-Bayesian Players. *International Journal of Game Theory* 9:157-67.

Miettinen T. 2007. Learning Foundation for the Cursed Equilibrium. mimeo.

Moore L, Juh S. 2006. Derivative Pricing 60 Years before Black-Scholes: Evidence from the Johannesburg Stock Exchange. *The Journal of Finance* 61:3069–98.

Monderer D, Samet S, Sela A. 1997. Belief Affirming in Learning Processes. *Journal of Economic Theory* 73: 438-52.

Monderer D, Shapley, L. 1996 . Potential Games. *Games and Economic Behavior* 14:124-43.

Murphy AH, Winkler R. 1977. Can Weather Forecasters formulate reliable forecasts of precipitation and temperature? *National Weather Digest* 2: 2-9.

Nachbar, J. 1997, Prediction, Optimization, and Learning in Repeated Games, *Econometrica*, 65: 275-309.

Noldeke G, Samuelson L. 1993. An Evolutionary Analysis of Forward and Backward Induction. *Games and Economic Behavior* 5: 425-54.

Norman M. 1968. Some Convergence Theorems for Stochastic Learning Models with Distance-Diminishing Operators. *Journal of Mathematical Psychology* 5: 61-101.

Nowak M, Sasaki A, Taylor C, Fudenberg D. 2004. Emergence of Cooperation and Evolutionary Stability in Finite Populations. *Nature* 428: 646-50.

Nowak M, Sigmund K. 1998a Evolution of indirect reciprocity by image scoring. *Nature* 393: 573-77.

Nowak M, Sigmund K. 1998b. The dynamics of indirect reciprocity. *Journal of Theoretical Biology* 194: 561-74.

Olszewski W, Sandroni A. 2006., Strategic Manipulation of Empirical Tests. Northwestern University mimeo.

Posch M, Sigmund K. 1999. The Efficiency of Adapting Aspiration Levels. *Proceedings of the Royal Society: Biological Sciences* 266:1427-14

Primiceri GE. 2006.Why Inflation Rose and Fell: Policy-Makers' Beliefs and US Postwar Stabilization Policy. *Quarterly Journal of Economics* 121, 867-901.

Rubinstein A. 1989. The electronic mail game: Strategic Behavior under ‘almost common knowledge. *American Economic Review* 79:335-91.

Rubinstein A, Wolinsky A. 1994. Rationalizable Conjectural Equilibrium: Between Nash and Rationalizability. *Games and Economic Behavior* 6: 299-311.

Saari D, Simon C. 1978. Effective Price Mechanisms. *Econometrica* 46: 1097-1125.

Salmon T. 2001. An Evaluation of Econometric Models of Adaptive Learning. *Econometrica* 69: 1597-1628.

Sandholm W. 2005. Excess payoff dynamics and other well-behaved evolutionary dynamics. *Journal of Economic Theory* 124:149-70.

Sandholm W. 2007 Evolution in Bayesian Games II: Stability of Purified Equilibria. *Journal of Economic Theory* 136:641-667

Sandholm W. 2009 *Population Games and Evolutionary Dynamics,* MIT Press. Cambridge, MA.

Sandroni A. 2003.The Reproducible Properties of Correct Forecasts. *International Journal of Game Theory* 32:151-159.

Sargent TJ. 2008. "Evolution and Intelligent Design," *AEA Presidential Address*.

Sargent TJ. 1999. *The Conquest of American Inflation*, Princeton:Princeton University Press.

Sargent TJ, Velde F. 2002. *The Big Problem of Small Change*, Princeton: Princeton University Press.

Savage L. 1954 *The Foundations of Statistics* New York, John Wiley and Sons.

Shamma JS, Arslan G. 2005. Dynamic fictitious play, dynamic gradient play, and distributed convergence to Nash equilibria. *IEEE Transactions on Automatic Control*, 50: 312-327.

Shapley L 1964 Some Topics in Two-Person Games. *Advances in Game Theory: Annals of Mathematical Studies* 5:1-28.

Stinchcombe M 2005 The Unbearable Flightiness of Bayesians: Generically Erratic Updating. mimeo.

Vovk V. 1990. Aggregating Strategies. *Proceedings of the 3rd Annual Conference on Computational Learning Theory*, 371-383.

Weber R 2003 Learning' with no feedback in a competitive guessing game. *Games and Economic Behavior* 44:134-44.

Wilcox N 2006 Theories of Learning in Games and Heterogeneity Bias. *Econometrica* 74: 1271-92.

Williams N. 2004. Stability and Long Run Equilibrium in Stochastic Fictitious Play. mimeo.

Young P. 2008. Learning by Trial and Error. Mimeo.

RELATED RESOURCES

Cressman R. 2003 *Evolutionary Dynamics and Extensive-Form Games.* MIT Press. Cambridge MA.

Fudenberg D, Levine DK 1998 *The Theory of Learning in Games.* MIT Press Cambridge, MA.

Hart S. 2005 Adaptive Heuristics. *Econometrica* 73:1401-1430.

Samuelson L. 1997 *Evolutionary Games and Equilibrium Selection.* MIT Press. Cambridge MA.

Young P. 2004 *Strategic Learning and Its Limits.* Oxford University Press, Oxford UK.